FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation

Year: 2023 | Citations: 282 | Authors: Tu Vu, Mohit Iyyer, Xuezhi Wang

---

Abstract

Most large language models (LLMs) are trained once and never updated; thus, they lack the ability to dynamically adapt to our ever-changing world. In this work, we perform a detailed study of the factuality of LLM-generated text in the context of answering questions that test current world knowledge. Specifically, we introduce FreshQA, a novel dynamic QA benchmark encompassing a diverse range of question and answer types, including questions that require fast-changing world knowledge as well as questions with false premises that need to be debunked. We benchmark a diverse array of both closed and open-source LLMs under a two-mode evaluation procedure that allows us to measure both correctness and hallucination. Through human evaluations involving more than 50K judgments, we shed light on limitations of these models and demonstrate significant room for improvement: for instance, all models (regardless of model size) struggle on questions that involve fast-changing knowledge and false premises. Motivated by these results, we present FreshPrompt, a simple few-shot prompting method that substantially boosts the performance of an LLM on FreshQA by incorporating relevant and up-to-date information retrieved from a search engine into the prompt. Our experiments show that FreshPrompt outperforms both competing search engine-augmented prompting methods such as Self-Ask (Press et al., 2022) as well as commercial systems such as Perplexity.AI. Further analysis of FreshPrompt reveals that both the number of retrieved evidences and their order play a key role in influencing the correctness of LLM-generated answers. Additionally, instructing the LLM to generate concise and direct answers helps reduce hallucination compared to encouraging more verbose answers. To facilitate future work, we release FreshQA at github.com/freshllms/freshqa and commit to updating it at regular intervals.